Kinetics and Reaction Coordinates of the Reassembly of Protein Fragments Via Forward Flux Sampling

Ernesto E. Borrero, Lydia M. Contreras Martínez, Matthew P. DeLisa, and Fernando A. Escobedo*
School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, New York

ABSTRACT We studied the mechanism of the reassembly and folding process of two fragments of a split lattice protein by using forward flux sampling (FFS). Our results confirmed previous thermodynamics and kinetics analyses that suggested that the disruption of the critical core (of an unsplit protein that folds by a nucleation mechanism) plays a key role in the reassembly mechanism of the split system. For several split systems derived from a parent 48-mer model, we estimated the reaction coordinates in terms of collective variables by using the FFS least-square estimation method and found that the reassembly transition is best described by a combination of the total number of native contacts, the number of interchain native contacts, and the total conformational energy of the split system. We also analyzed the transition path ensemble obtained from FFS simulations using the estimated reaction coordinates as order parameters to identify the microscopic features that differentiate the reassembly of the different split systems studied. We found that in the fastest folding split system, a balanced distribution of the original-core amino acids (of the unsplit system) between protein fragments propitiates interchain interactions at early stages of the folding process. Only this system exhibits a different reassembly mechanism from that of the unsplit protein, involving the formation of a different folding nucleus. In the slowest folding system, the concentration of the folding nucleus in one fragment causes its early prefolding, whereas the second fragment tends to remain as a detached random coil. We also show that the reassembly rate can be either increased or decreased by tuning interchain cooperativeness via the introduction of a single point mutation that either strengthens or weakens one of the native interchain contacts (prevalent in the transition state ensemble).

INTRODUCTION

Protein fragment complementation assays (PCAs) have been powerful experimental tools for assaying highly specific interactions involving cellular proteins (1–5). This approach is based on splitting a reporter protein into two individual fragments that by themselves remain inactive but on reassembly, yield the original properly folded and active protein structure. Examples of these systems include split green fluorescent protein (GFP) and its spectral variants YFP and CFP (2,6), ubiquitin (7), murine dihydrofolate reductase (mDHFR) (8), β-lactamase (9,10), and firefly luciferase (11). The reconstituted activity of all these proteins can be conveniently detected by fluorescence or well-established enzymatic assays. The application of PCAs in living cells has become an invaluable tool for mapping protein-protein and protein-nucleic acid interaction networks.

A major factor limiting the usefulness of split proteins is the slow folding kinetics and formation of misfolded aggregates that is associated with the reassembly process of multiple fragments (2,12). This inefficiency hinders the effective application of PCAs on biologically relevant timescales. For instance, although GFP fluorescence can be detected in minutes, the two fragments that result when the protein is dissected in the middle of the sequence fail to associate and reassemble when expressed in bacteria (2). Even when the fragments are each fused to strongly interacting leucine zippers (K_D ~1–2 mM), folding and fluorescence activity of the reconstituted protein is not observed until after 1–2 days (13). Given that similar drawbacks have been observed for different split reporters, significant effort has been focused on strategies to accelerate the formation of a reassembled protein. Some attempts include: i), identification of multiple permissive split sites along the protein, typically away from the catalytic site, using circular permutation (14,15); ii), structure-guided design, using bioinformatic and theoretical analysis (9,16,17); iii), optimization of target sequence using directed evolution for more efficient folding/reassembly (18,19); and, most recently, iv), the addition of hybridizing molecules (20). Despite these efforts, our understanding of how parameters such as the split site position in the primary sequence and size of resulting fragments contribute to the efficiency of protein reassembly remains limited.

Proteins can fold by diverse pathways including nucleation-condensation, framework (hierarchical) model, and hydrophobic collapse models (21). The nucleation-condensation mechanism describes the overall features of folding of most domains by uniting features of the other two folding models and invoking the formation of hydrophobic and long range interactions in the transition state (TS) to stabilize weak secondary structures. Given that in a nucleation folding mechanism a few key residues known as the folding nucleus provide a significant driving force in the formation of the TS leading to the folding of many proteins (22–24), a clear understanding of how these amino acids are distributed between fragments could be key to our progress in designing

Submitted August 14, 2009, and accepted for publication December 15, 2009.

*Correspondence: fe13@cornell.edu

Editor: Costas D. Maranas.

0006-3495/10/05/1911/10 $2.00

doi: 10.1016/j.bpj.2009.12.4329

efficient split protein systems. Intrigued by this notion, we previously used brute-force Monte Carlo simulations to analyze how the thermodynamics and kinetics of the reassembly process for two split protein model systems were affected by the location of the split site (25). In that study, we rationalized thermodynamically why reassembly of a split fragment system is significantly slower than the folding of an unsplit protein. We showed that strategic splitting of the folding nucleus, where the nucleus is more equally shared between the two fragments, drastically accelerated reassembly by: i), preventing the permanent preassembly of an individual fragment that would otherwise lead to a slower two-step assembly process where chain preassembly precedes interchain contacts; and ii), driving the formation of interchain native contacts that promoted a cooperative and productive folding. Interestingly, reassembly of split ubiquitin is observed experimentally when the protein is fragmented such that the amino acid residues that make up the compact hydrophobic core (26,27) have a 60–40% distribution between fragments (7). In Contreras Martínez et al. (25), however, we did not provide a precise characterization of the folding mechanism or of the TS.

Several transition path sampling methods have been developed to study the kinetics of biomolecular rare events (28,29) by enhancing the sampling of the transition region. These methods allow the estimation of rate constants and the collection of pathways (i.e., the transition path ensemble (TPE) to establish the transition mechanism, which would be impractical via conventional brute-force simulations. The technique of choice for this study is the forward flux sampling (FFS) method because of its simplicity and efficiency. Two of us (E. Borrero and F. Escobedo) used FFS previously to evaluate the kinetics of the transition pathways for the folding mechanism of a single chain (unsplit) 48-mer lattice protein (30). We showed that the initial formation of a critical core of amino acids (24,31) is the most important step during folding, a result that is relevant for proteins after a nucleation folding mechanism.

In this study, we further investigate the mechanism of reassembly of the same model 48-mer system and split systems, taking advantage of optimized FFS-type approaches (32,33) to analyze mechanistic details of their folding pathways (TPE) and estimate a suitable reaction-coordinate (RxC). The RxC essentially corresponds to the committor probability (p_B) surface, which gives the probability of any particular configuration to reach the final folded state. After the RxC is parameterized in terms of physically meaningful properties that describe the system's dynamics, the mechanistic details of the transition can be extracted by screening the microscopic properties of the ensemble of configurations belonging to different p_B isosurfaces (e.g., for $p_B = 0.5$, which corresponds to the TS). Our results provide insight to two interesting questions that arose from our previous analysis: i), whether the split fragments exhibited the same folding nucleus and nucleation-driving folding mechanism as the parent protein; and ii), whether the same folding mechanism for the parent protein (or a highly similar one) would be optimal for fragment reassembly.

SIMULATION DETAILS

In FFS, interfaces are used to partition the phase space along an order parameter λ connecting the initial and final regions of interest. On each point at each interface, multiple trial runs (k_i) are carried out to promote successful partial paths between interfaces. In the Supporting Material, a short description is given of the branched growth (BG) method adopted that can optimize the selection of λ as RxC, the spacing of λ, and the number of trial trajectories per point k_i.

Unsplit system

The 48-mer protein model adopted here exhibits a fast and stable proteinlike folding into a unique native structure via a two-state (unfolded-folded) process whose transition pathways are known to follow a nucleation-driven folding mechanism (30). The formation of a critical core of amino acids mediates the folding of the single-chain (unsplit) protein (23,24). This critical core is formed at an early stage of the process by those residues that have a higher chance of being in contact in the TS (30) and was composed of several (mostly hydrophobic) amino acid residues, that have >80% probability of forming native contacts (i.e., residues: 13, 16, 17, 19–24, 26–31, and 34–37). Fig. 1 shows the contact map density for all the native contacts belonging to the ensemble of configurations at isocommittor $p_B = 0.2$, 0.5, and 0.8 surfaces. The configurations belonging to different regions of the isocommittor surface were collected by calculating the p_B value for each TPE interfacial configuration via the following equation:

$$p_B(NC, E) = -0.404 + 0.017(NC) - 0.029(E). \quad (1)$$

This RxC model estimates the probability of any TPE interfacial point to commit to the folded state from the *NC* and *E* values of that point (30). Fig. 1, *center*, shows that the 15 critical core (CC) native contacts with higher probability to belong to the TS ($p_B = 0.5$) are formed by those residues forming the folding nucleus. Fig. 1, *top*, shows that these CC native contacts start to form the nucleus at early stages of the folding process ($p_B = 0.2$). Fig. 1, *bottom*, shows that at late stages of the folding process ($p_B = 0.8$), the protein acquires its native structure by forming contacts around the folding nucleus.

Split system preparation

The split lattice model proteins were generated by dissecting the 48-mer at one of three possible positions: between residues 16 and 17 (N-split), 24 and 25 (M-split), and 32 and 33 (C-split). In all these cases the folded state is identical to the one reached by the single 48-mer chain and

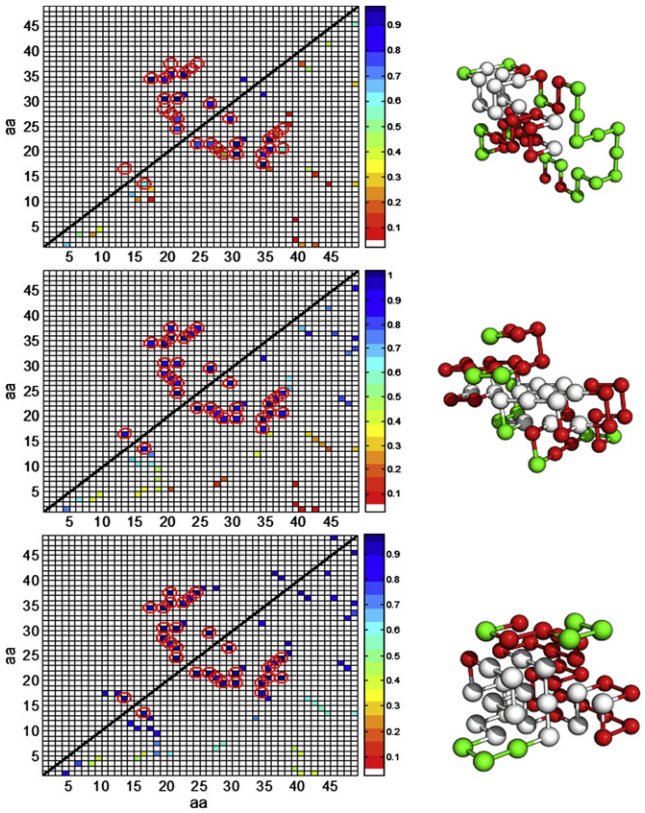

FIGURE 1 Contact density map for the unsplit 48-mer system for ensembles of configurations belonging to isocommittor surfaces: $p_B = 0.2$ (*top*), $p_B = 0.5$ (*center*), and $p_B = 0.8$ (*bottom*). The x and y axis represent the amino acid (aa) position in the 48-mer sequence. The ensembles were collected by estimating p_B values for all the interfacial points in the TPE from the RxC model in Eq. 1. The lower triangle (below *the diagonal line*) shows the probability of a native contact to belong to the ensemble; the color code is given by the vertical bar. The lower triangle at the isocommittor surface $p_B = 0.8$ (*bottom*) shows the 57 pair contacts for the native structure. The upper triangle shows those native contacts with at least 80% probability to belong to the corresponding p_B ensemble. Encircled symbols represent native contacts that form the critical folding nucleus. Snapshots depicting typical configurations observed for each ensemble are also shown where red/dark gray indicates native contacts and white indicates native contacts that form the critical folding nuclei.

characterized by NNC =58 native contacts. Table 1 gives the main characteristics for all the systems (i.e., the split and the unsplit systems), including the number of interchain CC contacts (InterC) that involve interacting residues from both chains and intrachain CC contacts (IntraC) that involve interacting residues within the same chain. These InterC and IntraC are formed by the original 15 critical core native contacts (identified in Fig. 1) on protein fragmentation. Note that in the C- and Mid-split cases, the folding core residues are well distributed between fragments and give a significant number of InterC. In contrast, for the N-split system most of the folding core residues are concentrated in chain B and are not involved in InterC (25). Moreover, the N-split and the C-split systems are symmetrical, with each system having one 16-mer fragment and one 32-mer fragment; this allows a comparison in the absence of chain length disparities. Further details on the model unsplit and split systems, including their structure and thermodynamics are given in Contreras Martínez et al. (25).

Conformational sampling

Conformational local sampling was carried out through a set of MC moves based on the Verdier-Stockmayer algorithm (34). Relative to these local moves, whole-fragment diffusional translation of a randomly selected chain was also attempted after each MC step with a priori probability ($\leq 10^{-4}$) (25). For simulating the folding kinetics, the temperature was fixed at $T = 0.25$, a value close to the folding transition temperature of the unsplit system. The system was confined inside a relatively large 3-D cubic box of side length (L) 12 σ (where σ is the lattice size = size of a protein residue) corresponding to a protein volume fraction of ~3% (25). Because of this dilution, it is assumed that spatial restriction affects the translational entropy of symmetrically and asymmetrically split systems in a commensurate way, and has negligible effect on conformational entropy (25,30). In comparing different split protein systems, the analysis in our previous work (25) indicated that differences in thermodynamic and kinetic behavior were not determined by diffusion limitations of the fragments trying to find each other. The spatial constriction also mimics a moderately crowded environment relative to open space, ensuring a timely association of the different fragments.

Candidate collective variables

In the simulations, the following macroscopic properties were calculated for all the state points collected at the λ interfaces in the TPE trajectories: total number of native contacts (NC), number native contacts in chain A (NNA), number native contacts in chain B (NNB), number of contacts between fragments (IC), number of native contacts between fragments (INC), conformational energy (E), and the number of critical core contacts (CC) (the latter as identified for the unsplit system). These collective variables were used for the RxC analysis via the FFS-LSE method.

RESULTS

A first preliminary BG simulation was carried out using the number of native contacts as initial guess of the order parameter (i.e., λ = NC) with the purpose of optimizing the position (λ values) and sampling (k values) of 12 interface ensembles. Details of this calculation and its results are given in the Supporting Material. These optimized parameters were then used to obtain the p_B history data via BG simulations with the FFS-LSE method (see Supporting Material). These p_B data were then used to screen a set of candidate collective properties (see above) for an optimized order parameter model λ, as described in the Supporting Material. Thereafter,

TABLE 1 Characteristics of the 48-mer and split systems

	Residues (n)				Native contacts			Folding nucleus contacts		
									IntraC	
	Chain A	chain B	NC	Energy [k_BT]	NNA	NNB	INC	InterC	Chain A	Chain B
48-1 mer	48	—	57	−20.24	57	—	—	—	15	—
C-split	32	16	58	−20.62	25	7	26	7	8	—
M-split	24	24	58	−20.65	16	12	30	13	2	—
N-split	16	32	58	−20.43	9	28	21	—	1	14

NNA and NNB are the number of intrachain contacts formed in chain A and B, respectively. INC is the number of native contacts formed between the two fragments. The distribution of the 15 native contacts forming the folding nuclei in split-proteins: interchain contacts (InterC) and intrachain contacts (IntraC).

a better estimate for the order parameter is obtained and used to partition the phase space for additional FFS-LSE simulations. The adaptive optimization algorithm is also applied to find a better λ staging of the new order parameter between iterations. This combination of FFS-LSE and staging optimization provides the advantage of allowing a more efficient and uniform distribution of the p_B data over the entire phase space, which is important to construct suitable RxC models, as discussed in Borrero and Escobedo (33). The combined scheme is repeated until similar RxC models are obtained in consecutive iterations.

Table 2 shows the coefficients for the RxC models obtained from the iterations of the combined scheme. For the first iteration, the LSE and ANOVA for model 1 in Table 2 indicates that the number of native contact between fragments (INC) is the only and most significant collective variable that describes the system transition to the folded state. Note that the number of native contacts (NC) is a global property that is too blunt to correlate with the number of important contacts for reassembly. A second iteration was then carried out using this new estimate of RxC. Again, the model surfaces were initially fitted to the set of candidate collective variables (see above) and ANOVA predicted significant linear, quadratic and interaction terms for the variables INC, NC, and E. A second LSE fitting was then carried out using only these three collective variables to obtain the RxC model 2 in Table 2. Additional iterations of the FFS-LSE algorithm converge to similar estimates for the RxC as model 2. The initial and optimized $\{\lambda_i'\}$ sets for model 2 of all split systems are given in Table S2. The p_B surfaces predicted by the RxC of these models 2 are illustrated in Fig. S1 and Fig. S2, indicating that in addition to variables E and NC, variable INC, interaction and quadratic terms are necessary for a more complete description of the isocommittor surface curvature. The dependence of the predicted TS isocommittor surface ($\lambda = p_B = 0.5$) on the INC variable, underlining the important role played by the interchain interactions in the transition, is consistent with the behavior observed by the mechanistic analysis discussed in the rest of the study. Fig. S3, Fig. S4, and Fig. S5 illustrate the qualitative differences in how p_B is correlated by the significant and the nonsignificant variables in the model.

The folding rates for the N-, Mid- and C-split systems are ~12, 19, and 31% the rate of the 48-mer, respectively. More revealing, the N-split system folds at a rate that is 37% that of the C-split system (despite having fragments of equal lengths) and 61% that of the Mid-split system (Fig. S6).

TABLE 2 FFS-LSE parameters for the reaction coordinate (RxC) model [$p_B \approx \beta + \beta_1 NC + \beta_2 INC + \beta_3 E + \beta_4 NC^2 + \beta_5 INC^2 + \beta_6 E^2 + \beta_7 NC \times INC + \beta_8 NC \times E + \beta_9 INC \times E$]

		Model coefficient (β) [F_0]									
System	Model	(β_0)	NC	INC	E	NC^2	INC^2	E^2	NC × INC	NC × E	INC × E
N-split	1	−0.75	—	0.083 [14913]	—	—	—	—	—	—	—
	2	−1.78	0.069 [27]	−0.038 [108]	−0.095 [19]	−0.003 [147]	−0.003 [179]	−0.019 [119]	−0.002 [32]	−0.005 [40]	−0.01 [90]
Mid-split	1	−0.43	—	0.048 [1707]	—	—	—	—	—	—	—
	2	−0.86	0.019 [16]	0.018 [53]	−0.043 [158]	−0.001 [30]	−0.005 [333]	−0.002 [166]	0.005 [538]	—	—
C-split	1	−0.82	—	0.079 [2020]	—	—	—	—	—	—	—
	2	−1.87	0.060 [13]	0.016 [53]	−0.084 [12]	−0.002 [82]	−0.009 [777]	−0.010 [22]	0.005 [140]	−0.006 [28]	−0.003 [17]

The collective variables are defined in the text. The significance of any individual regression coefficient for the model description is indicated by the partial $F_0 = MS_{SSR}/MS_E$ value; i.e., the ratio of the regression sum-square due to β_j and the mean-square for the residuals. For any nonzero regression term, the p-value for the F_0 statistics is $<\alpha$ (here chosen to be to 0.05 for a 95% confidence interval). The β_j significance increases with the F_0 value.

In our previous study, these differences in folding kinetics were rationalized by comparing the differences in free energy barriers observed between the 48-mer and the different split systems. We found a shift in the TS dividing surface especially for the N-split and Mid-split systems relative to the 48-mer transition, suggesting that the reassembly of these systems takes place via a different, slower folding mechanism. An alternative explanation that we investigate in this work is that the faster C-split system uses an alternative mechanism that is more efficient for assembly.

Perusing the available RxCs, we can re-examine the thermodynamics of the split systems to search for clues that may explain the differences observed in their kinetic behavior. In particular, we define free energy (A), energy (E), and entropy (S) changes between the unfolded and folded states (ΔA, ΔE, and ΔS) and between the TS and the unfolded states (ΔA^*, ΔE^*, and ΔS^*); e.g., ΔA and ΔA^* are estimated from

$$\Delta A = A^{\text{unfolded}} - A^{\text{folded}} = -k_B T \ln\left(\frac{P_u}{P_f}\right), \quad (2)$$

$$\Delta A^* = A^{TS} - A^{\text{unfolded}} = -k_B T \ln\left(\frac{P_{TS}}{P_u}\right), \quad (3)$$

where P_f, P_u, and P_{TS} are the probabilities of the folded, unfolded, and TS, respectively. Note that the optimized λ(INC,NC,E) in Table 2 are only suitable to estimate p_B values for configurations constrained in the phase space between the two stable states, i.e., configurations with $\lambda \leq 0$ and $\lambda \geq 1$ values are assigned to the $p_B = 0$ and $p_B = 1$ ensembles, respectively (32). P_{TS} was obtained by encompassing the free energy maximum close to the ensemble of configurations with $p_B = 0.5$. The energy change associated with Eq. 2 or Eq. 3 can be computed from the difference between the average configurational energy of the folded state (or TS) and unfolded state. The entropic contributions can be estimated from $T\Delta S = \Delta E - \Delta A$ and $T\Delta S^* = \Delta E^* - \Delta A^*$. For all the split systems, the changes in A, E, and S as the system goes from the unfolded to the TS and to folded states are depicted in Fig. 2.

The reason why we calculate differences with respect to the folded state in Eq. 2 is because the energy and entropy (and thus the free energy) of the ''unique'' folded configuration is approximately the same for all systems (e.g., $E^{\text{folded}} \sim -20.5\ k_B T$, $S^{\text{folded}} \approx 0$); consequently, comparing ΔA, ΔE, and ΔS across systems highlights the differences in the unfolded states. Although such differences in unfolded states are likely very large in configuration space, Fig. 2 shows that they are rather small in the A, E, and S spaces ($<kT$), once the unfolded state is standardized (or renormalized) according to our $p_B = 0$ model definition; indeed, all split systems show a similar transitional behavior. Another general feature in Fig. 2 is that the entropy change between the TS and the folded state is minimal for all systems, indicating that most of the translational and conformational freedom has been lost by the time the TS is reached (recall that $S^{\text{folded}} \approx 0$). In the context of TS theory, where the rate constant is $k_{A\to B} \approx C \exp(-\Delta A^*/kT)$, the similarity of ΔA^* for all split systems suggests that thermodynamic quantities alone will not be strongly indicative of kinetic behavior. Instead, differences in the frequency factor C are likely significant (keeping in mind that TS theory itself has a limited interpretative value as suggested by Fig. S2). Nonetheless, some useful correlations can be established in light of the small but real differences in energetic and entropic changes.

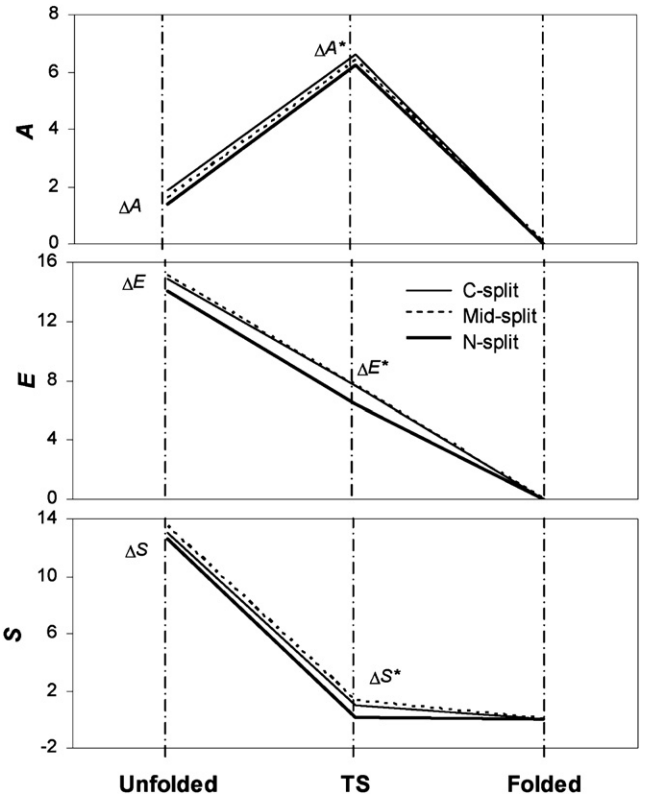

FIGURE 2 Schematic showing the difference in free energy, energy, and entropy for the different split systems as they go from the unfolded state, through the TS, and to the folded or reassembled state. For convenience, the properties for the folded state are all set to the same value of zero.

In the N-split case, the unfolded state is characterized by conformations having no interactions between the two fragments, open conformations of the small chain A, and compact, prefolded conformations for chain B. The C-split system exhibits a more cooperative folding behavior, where the unfolded state is characterized by configurations with both chains attached such that the total entropy of the system is essentially purely conformational. These characteristics of the unfolded state cause a nontrivial interplay of the interactions between the fragments. For example, given that the concentration of the folding core amino acids in a single fragment (chain B) stabilizes the unfolded state, the N-split system has stronger (more negative) energetic interactions than the unfolded C-split case ($E^{\text{unfolded}}_{\text{N-split}} - E^{\text{unfolded}}_{\text{C-split}} = -0.67kT$);

resulting in a smaller enthalpic driving force in going from unfolded to the folded state ($\Delta E_{N-split} < \Delta E_{C-split}$ in Fig. 2). In contrast, the total entropy drop (penalty) from the unfolded to the folded state is smaller, more favorable for the N-split; i.e., $\Delta S_{N-split} < \Delta S_{C-split}$. Although the drop in translational entropy is expected to be larger for the N-split case, the prefolding of the large fragment at early stages of the folding decreases its configurational entropy, resulting in a smaller overall entropy drop than for the other split systems (as shown in Fig. 2). Overall, the free energy drop from the unfolded to the folded state for the N-split is slightly smaller than for the other systems ($\Delta A_{N-split} < \Delta A_{M-split} < \Delta A_{C-split}$ in Fig. 2), indicating that the reassembly driving force is smaller for the N-split system. Henceforth, differences among systems are discussed in the context of the kinetic FFS data.

The mechanistic details for the reassembly transition of the split systems were obtained by collecting ensembles of configurations at different iso-lines of the committor surface $p_B = 0.2$, 0.5, and 0.8 (shown in Figs. 3–5). The configurations were classified during a FFS simulation using $\lambda = p_B$ model 2 (Table 2) as RxC. Each such ensemble was then analyzed at a microscopic level by determining the probability of each native contact pair to belong to the corresponding ensemble.

For the N-split system, Fig. 3 shows the contact density map and snapshots of typical configurations for the $p_B = 0.2$, 0.5, and 0.8 ensembles. Native interactions between the two fragments begin at late stages of the folding process as the TSE is characterized by structures in which interchain interactions constitute only 10% (2 of 20) of the most probable native contacts. Interestingly, 69% and 63% of those native contacts with a minimum 80% probability to belong to the $p_B = 0.2$ and $p_B = 0.5$ surfaces, respectively, correspond to the unsplit-system CC set (compare to Fig. 1). The nucleus formation starts in the chain B at early stages of the folding process ($p_B = 0.2$) by forming 73% of the CC set. Additionally, 80% of those most probable native contacts at the $p_B = 0.2$ and 0.5 surfaces are formed in chain B indicating that this chain prefolds, separate from chain A that remains as a random coil. It is only after the reaction pathway crosses the $p_B = 0.5$ TS dividing region, that the number of native contacts between chains and intrachain contacts in chain B increases to achieve the final folded state. However, only 2 of 30 of those most probable native contacts at the $p_B = 0.8$ iso-surface correspond to intrachain contacts in the small A fragment, indicating that chain A completes its folding at a very late stage of the reassembly process, when it associates with chain B. This kinetic analysis supports the picture we gathered before from a thermodynamic analysis (25) that argued that although the prefolding of chain B favors folding on entropic grounds (i.e., the unfolded and folded states have similar amount of order) it disfavors folding on energetic grounds. For the N-split system, this interplay of interactions causes retardation of the folding mechanism.

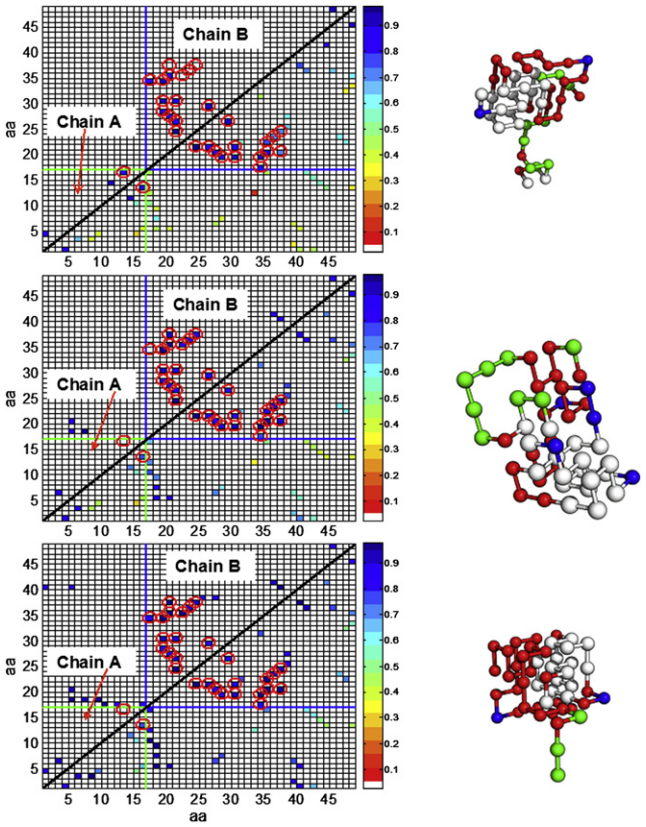

FIGURE 3 Contact density map for the N-split system for the ensembles of configurations belonging to isocommittor surfaces: $p_B = 0.2$ (*top*), $p_B = 0.5$ (*center*), and $p_B = 0.8$ (*bottom*). The x and y axis represent the amino acid (aa) position in the 48-mer sequence. The ensembles were collected by estimating p_B values for all the interfacial points in the TPE from RxC model 2 in Table 2. The lower triangle (below *the diagonal line*) shows the probability of a native contact to belong to the ensemble. The upper triangle shows those native contacts with at least 80% probability to belong to the corresponding p_B ensemble. Encircled symbols represent native contacts that form the original critical folding nucleus. Snapshots depicting typical configurations observed for each ensemble are also shown where green/light gray indicates chain A, blue/black indicates chain B, red/dark gray indicates native contacts, and white indicates native contacts that form the critical folding nuclei. The large squares in dashed lines enclose native intrachain contacts: (*green/light gray*) chain A and (*blue/black*) chain B.

In the case of the fast folding C-split system, native interactions between chains start at early stages of the folding process such that interchain associations constitute ~20% and 30% of the most probable native contacts at the $p_B = 0.2$ and $p_B = 0.5$ isosurfaces, respectively. Fig. 4 shows the contact density map and snapshots of typical configurations for the C-split system at the $p_B = 0.2$, 0.5, and 0.8 surfaces. CC contacts are only 31% and 35% of those native contacts with at least 70% probability to belong to the $p_B = 0.2$ and $p_B = 0.5$ ensembles, respectively. Moreover, the same nine most probable CC contacts formed at the TSE are also observed in the $p_B = 0.8$ ensemble. Surprisingly, we can infer that the nucleation in the C-split system uses a different set of native CC compared to the 48-mer protein

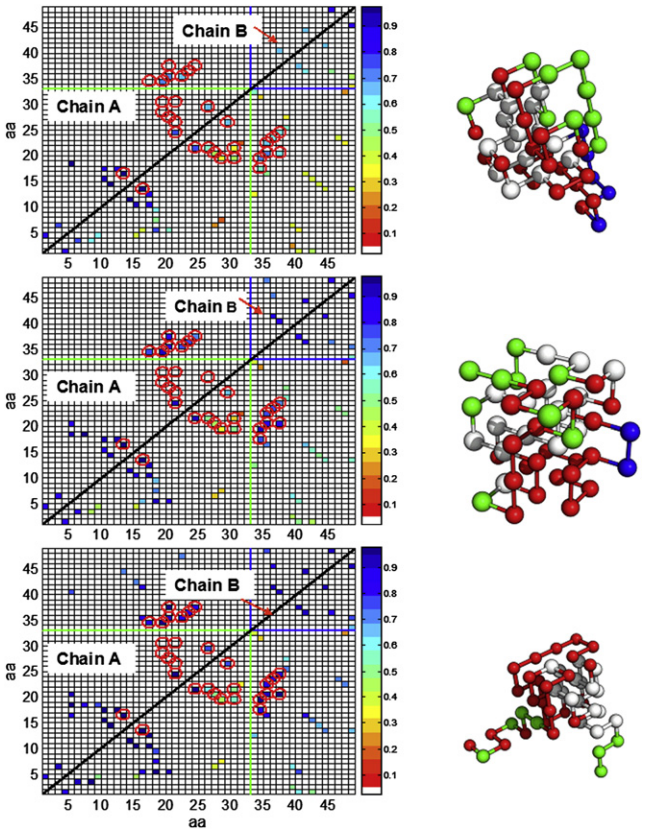

FIGURE 4 Contact density map for the C-split system for ensembles of configurations belonging to isocommittor surfaces: $p_B = 0.2$ (*top*), $p_B = 0.5$ (*center*), and $p_B = 0.8$ (*bottom*). The axes, symbols, lines, and colors have the same interpretation as given in the caption of Fig. 3, except that the upper triangle shows those native contacts with at least 70% probability to belong to the given p_B ensemble.

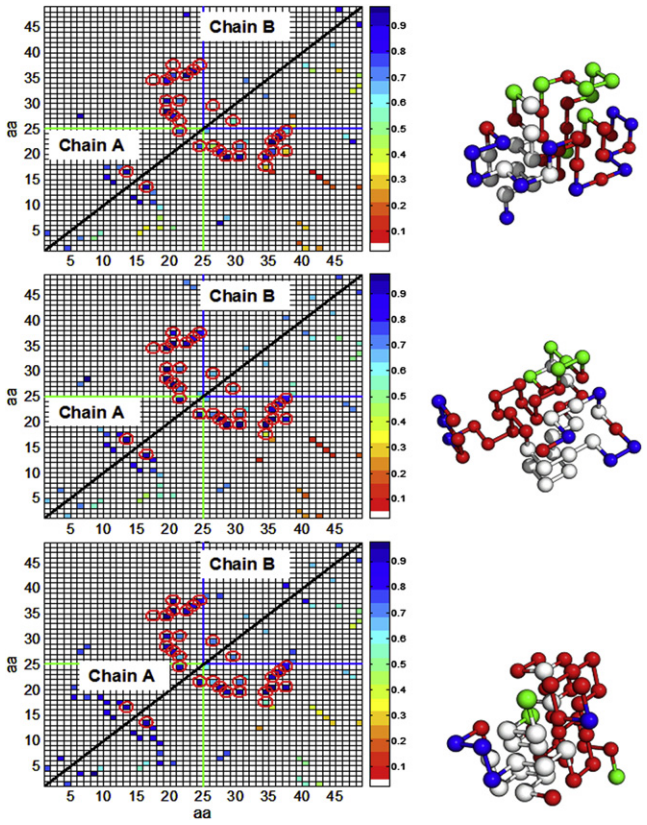

FIGURE 5 Contact density map for the Mid-split system for the ensembles of configurations belonging to isocommittor surfaces: $p_B = 0.2$ (*top*), $p_B = 0.5$ (*center*), and $p_B = 0.8$ (*bottom*). The axes, symbols, lines, and colors have the same interpretation as given in the caption of Fig. 3, except that the upper triangle shows those native contacts with at least 70% probability to belong to the given p_B ensemble.

(see Fig. 1), i.e., a different folding nucleus. This finding is unexpected given our previous conjecture (25) that the C-split and the 48-mer systems had a similar nucleation mechanism because they had the TS dividing surface located in the same position along the conformational energy as RxC (see Fig. 3 in Contreras Martínez et al. (25)). As in the unsplit system, the CC for the C-split system is also defined by those 15 residues that have a higher chance of being in contact in the TS. In this case the core is given by five interchain contact pairs (of 26 in folded state), seven intrachain A contact pairs (of 25 in folded state), and three intrachain B contact pairs (of seven in folded state). Moreover, seven of those 15 core pairs for the C-split correspond to CC contacts of the 48-mer protein and five of these seven pairs correspond to the most probable CC contacts in the unsplit TSE (see Table 2 in Contreras Martínez et al. (25)) that also form the A-B interchain contacts in the C-split TSE. Note that if the folding nuclei for the C-split were given by the same CC set of the 48-mer protein, all those native pairs would correspond to interchain contacts and intrachain contacts in chain A at the TSE. Fig. 4 also shows that the set of most probable native contacts at the TSE includes six of seven intrachain contacts in chain B of the C-split system, suggesting that this small chain is folded at that stage.

In our previous work (25) we hypothesized that the shift of the TS dividing region to the folded basin evidenced that the reassembly of the N-split system takes place via a different folding mechanism compared to the Mid- and C-split systems (see Fig. 3 in Contreras Martínez et al. (25)). However, our results in Fig. 3 suggest that the N-split folding mechanism entails a fast nucleation event in the large fragment (chain B) that is significantly similar to that of the folding nucleus (CC set) in the unsplit system. Of course that the similarity in the set of core contacts at the TS surface for the N-split and the unsplit 48-mer protein need not translate into a similarity in the rates of folding. These results for the C-split and N-split systems illustrate that any conclusions drawn based on shifts (or lack thereof) of the TS dividing surface are risky, especially if a poor order parameter is used as RxC.

For the Mid-split system, the contact map density and snapshots of typical configurations for the $p_B = 0.2$, 0.5, and 0.8 surface regions are shown in Fig. 5. Similar to the C-split system, the native interactions between chains begin at early stages of the folding process. Hence, interchain

associations correspond to ~56% and 67% of those native contacts with at least 70% to belong to the ensemble of configurations at $p_B = 0.2$ and TSE, respectively. However, there are two main differences between the folding mechanisms for the Mid- and C-split systems; in the former case the CC set provides a shared folding nucleus between chains that glues fragments together, and the folding of both chains is preceded by a cooperative association of the chains (i.e., they fold while attached one to another). Fig. 5 suggests that the transition of the Mid-split system to the folded state follows a nucleation mechanism similar to that observed in the unsplit system. Indeed, 67% and 93% of the CC set is formed by those native contacts most likely to occur in the $p_B = 0.2$ and $p_B = 0.5$ ensembles, respectively. In this case, a prefolded state of one of the chains is not observed at the TSE (in contrast to the C-split case), indicating that the formation of intrachain contacts in both chains is given by a cooperative folding behavior. The rate constant for the Mid-split transition is ~1.6 times that of the N-split transition, but 0.6 times that of the C-split transition. Note that the early prefolding of the small chain for the C-split decreases slightly its configurational entropy compared to the Mid-split case ($\Delta S_{C-split} < \Delta S_{Mid-split}$ in Fig. 2). The early strong interchain association for the Mid-split system balances out the early prefolding of the small chain in the C-split system to produce comparable energy changes (e.g., $\Delta E_{C-split} \approx \Delta E_{Mid-split}$); overall, the free-energy drop is slightly larger for the C-split system.

A common characteristic shared by the faster folding systems (Mid- and C-split) is that the formation of a few CC native contacts at an early stage facilitates the reassembly via a more cooperative behavior. We then hypothesized that if a single point mutation is introduced to the N-split case, resulting in a new mutN-split system, by strengthening one of the prevalent native interchain contacts observed in the TSE, then chains would associate earlier and accelerate reassembly. To test this idea, the mutN-split case incorporated a pseudo point-mutation with double the normal contact energy for residue 5 in chain A and residue 2 in chain B (this was the stronger interchain NC pair observed in the TSE for the N-split system as seen in Fig. 3). Note that this change does not involve any residue in the core nucleus. Fig. S7 shows the corresponding contact map density for the $p_B = 0.2$, 0.5, and 0.8 isosurfaces for this mutant system. As expected, the pseudo mutation causes interchain interactions at early stages of the folding process, corresponding to 25% and 40% of those native contacts most observed in the $p_B = 0.2$ and $p_B = 0.5$ ensembles, respectively. Moreover, Fig. S7 suggests that the reassembly of the mutN-split system still follows a nucleation mechanism similar to that observed in the N-split system, because now 27%, 40%, and 100% of the CC set is formed by those most probable native contacts in the $p_B = 0.2$, 0.5, and 0.8 ensembles, respectively. The fact that 50% and 78% of those most probable native contacts at the $p_B = 0.5$ and 0.8 surfaces, respectively, are formed in chain B, indicates that this chain prefolds. The average rate constant for the mutN-split transition is ~2 times greater than that of the unmutated N-split transition (see Fig. S6), but it remains ~20% slower than in the C-split system, suggesting that the early association of the chains may not be enough to achieve the reassembly efficiency exhibited by the C-split system.

Finally, to test that the distribution of the core residues in the C-split is key for the early interchain association and faster folding rate, we introduced to the C-split case a single point mutation by halving the contact energy for residue 19 in chain A and residue 2 in chain B. This is one of the strongest hydrophobic core pairs (see Table 2 in Contreras Martínez et al. (25)) and one of the prevalent native interchain contacts observed in the TSE (Fig. 4). The average rate constant for this mutated C-split transition is over 4 times smaller than that of the unmutated C-split transition. Consistent with our expectation, this is due to the disruption of the nucleation mechanism observed in Fig. 4 (i.e., a different core nucleus as shown in Fig. S8).

CONCLUSIONS

In this work, we used FFS to study the reassembly mechanism of a model 48-mer system. Our results support our previous findings (25) that reassembly of a split fragment system was significantly retarded relative to the unsplit parental protein, and that the nature and magnitude of reassembly retardation was highly dependent on the distribution of the critical nuclei between the two split fragments. We observed that the most efficient folding system (the C-split system) has a more balanced distribution of core amino acid residues between the two fragments (52–38%) relative to the other slower systems (31–69% for the Mid-split and 7–93% for the N-slit system) that promotes interchain interactions at early stages in the reassembly process. The notion that the shared folding nucleus was critical to the formation of early interactions between chains that lead to productive folding was supported by the slow reassembly of the N-split system, where the critical core was localized in a single fragment. It is interesting to note that the reassembly of the split ubiquitin protein is observed experimentally when the protein is fragmented such that the amino acid residues making up the compact hydrophobic core have a 60–40% distribution between fragments (7).

The early interchain association that takes place in the fast-folding (C-split) system promotes cooperative folding (coassembly) between the two fragments, where they fold simultaneously into the final structure. In contrast, the concentration of the folding core in a single fragment (as in the case of the N-split system) leads to a two-step assembly process, where an individual fragment permanently preassembles and then forms connections with the second chain to reconstitute the native structure. The observation of coassembly for the fastest folding system (C-split) system, the most similar

to the parent protein in terms of folding rate, and a two-step folding for the slowest folding (N-split) protein led us to hypothesize that the concentration of core native contacts in a single fragment changed the cooperative folding mechanism observed in the parent protein (where all amino acids are linked) to a coassembly process. Intuitively, it seemed that sharing the core between two fragments preserved the overall folding mechanism exhibited by the parent protein so that the process was still productive when the protein is fragmented. However, a precise characterization of the folding mechanism and the TS for the split systems revealed the surprising result that the folding mechanism of the unsplit protein was unchanged in the slower folding (N-split and Mid-split) systems but significantly changed in the case of the fast folding (C-split) system, where a different TS was observed.

Given the higher degree of freedom of split protein systems (as compared to a single protein chain) to search for multiple folding pathways, it is significant that a split protein system can reassemble via the same folding mechanism as its parent structure. Yet, it is possible that the early commitment to the formation of the same parental TS and folding pathway could be a suboptimal strategy to reassemble the split system (as in the N-split case), because the system might not fully explore alternative, faster pathways. We therefore speculate that the utilization of a folding pathway different from the one used by the unsplit parent protein, leads to the efficient reassembly of the C-split fragment system. It remains unclear, however, whether optimal reassembly via a novel folding pathway is a general phenomenon or a highly system-specific occurrence in our split protein systems. Also, we cannot at present rule out the existence of a split point yielding a system that folds as fast as the C-split system and follows the mechanism of the parent protein.

One of the experimental strategies to control the folding or assembly mechanism of a protein system involves changes to amino acids and protein domains; e.g., the addition of leucine zipper domains has been experimentally observed to enhance fragment interactions and promote fragment reassembly for several split protein systems such as GFP, DHFR and ubiquitin (2,6–12). In this study, the rational introduction of a single point mutation to the N-split system was found to promote fragment association and speed up the reassembly kinetics. This mutation, however, did not change the overall folding mechanism nor did it match the folding efficiency of the C-split system, suggesting that additional factors to early interchain interactions contributed to the efficient folding of the latter. Although our results for a minimalist lattice model are not directly applicable to real split proteins, they suggest that reassembly of split protein fragments could be optimized by designing strategic fragmentation patterns that lead to different, more efficient folding mechanisms or by altering the sequence itself in a manner that promotes the interaction between fragments without necessarily affecting the overall folding process. The use of such key mutations presents an alternative to the introduction of additional protein domains (i.e., leucine zippers) that have been reported to be important to the natural activity and mode of action of several well-characterized proteins (35). With respect to experiments, we plan to test some of these predictions in the split ubiquitin system model by introducing various mutations that affect its well-characterized compact hydrophobic core. With respect to simulations, current work is directed to the reassembly mechanism of split lattice protein models exhibiting two domains (corresponding to independently secondary motifs) and following a hierarchical folding mechanism; future work will also aim at studying the kinetics of small (computationally tractable) atomistic split-protein models.

SUPPORTING MATERIAL

Two tables and eight figures are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(10)00133-5.

This work was supported by the National Science Foundation (grant 0756248).

REFERENCES

1. Hu, C. D., and T. K. Kerppola. 2003. Simultaneous visualization of multiple protein interactions in living cells using multicolor fluorescence complementation analysis. *Nat. Biotechnol.* 21:539–545.
2. Magliery, T. J., C. G. M. Wilson, …, L. Regan. 2005. Detecting protein-protein interactions with a green fluorescent protein fragment reassembly trap: scope and mechanism. *J. Am. Chem. Soc.* 127:146–157.
3. Ozawa, T., Y. Sako, …, Y. Umezawa. 2003. A genetic approach to identifying mitochondrial proteins. *Nat. Biotechnol.* 21:287–293.
4. Remy, I., and S. W. Michnick. 2004. A cDNA library functional screening strategy based on fluorescent protein complementation assays to identify novel components of signaling pathways. *Methods.* 32:381–388.
5. Stains, C. I., J. R. Porter, …, I. Ghosh. 2005. DNA sequence-enabled reassembly of the green fluorescent protein. *J. Am. Chem. Soc.* 127: 10782–10783.
6. Ghosh, I., A. D. Hamilton, and L. Regan. 2000. Antiparallel leucine zipper-directed protein reassembly: application to the green fluorescent protein. *J. Am. Chem. Soc.* 122:5658–5659.
7. Johnsson, N., and A. Varshavsky. 1994. Split ubiquitin as a sensor of protein interactions in vivo. *Proc. Natl. Acad. Sci. USA.* 91:10340–10344.
8. Pelletier, J. N., F. X. Campbell-Valois, and S. W. Michnick. 1998. Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments. *Proc. Natl. Acad. Sci. USA.* 95:12141–12146.
9. Galarneau, A., M. Primeau, …, S. W. Michnick. 2002. Beta-lactamase protein fragment complementation assays as in vivo and in vitro sensors of protein protein interactions. *Nat. Biotechnol.* 20:619–622.
10. Wehrman, T., B. Kleaveland, …, H. M. Blau. 2002. Protein-protein interactions monitored in mammalian cells via complementation of beta -lactamase enzyme fragments. *Proc. Natl. Acad. Sci. USA.* 99: 3469–3474.
11. Paulmurugan, R., and S. S. Gambhir. 2003. Monitoring protein-protein interactions using split synthetic *Renilla* luciferase protein-fragment-assisted complementation. *Anal. Chem.* 75:1584–1589.
12. Deo, S. K. 2004. Exploring bioanalytical applications of assisted protein reassembly. *Anal. Bioanal. Chem.* 379:383–390.

13. Wilson, C. G., T. J. Magliery, and L. Regan. 2004. Detecting protein-protein interactions with GFP-fragment reassembly. *Nat. Methods.* 1:255–262.

14. Hennecke, J., P. Sebbel, and R. Glockshuber. 1999. Random circular permutation of DsbA reveals segments that are essential for protein folding and stability. *J. Mol. Biol.* 286:1197–1215.

15. Iwakura, M., T. Nakamura, …, K. Maki. 2000. Systematic circular permutation of an entire protein reveals essential folding elements. *Nat. Struct. Biol.* 7:580–585.

16. Betton, J. M., and M. Hofnung. 1994. In vivo assembly of active maltose binding protein from independently exported protein fragments. *EMBO J.* 13:1226–1234.

17. Paszkiewicz, K. H., M. J. E. Sternberg, and M. Lappe. 2006. Prediction of viable circular permutants using a graph theoretic approach. *Bioinformatics.* 22:1353–1358.

18. Cabantous, S., T. C. Terwilliger, and G. S. Waldo. 2005. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat. Biotechnol.* 23:102–107.

19. Paulmurugan, R., and S. S. Gambhir. 2007. Combinatorial library screening for developing an improved split-firefly luciferase fragment-assisted complementation system for studying protein-protein interactions. *Anal. Chem.* 79:2346–2353.

20. Demidov, V. V., N. V. Dokholyan, …, N. E. Broude. 2006. Fast complementation of split fluorescent protein triggered by DNA hybridization. *Proc. Natl. Acad. Sci. USA.* 103:2052–2056.

21. Daggett, V., and A. R. Fersht. 2003. Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.* 28:18–25.

22. Abkevich, V. I., A. M. Gutin, and E. I. Shakhnovich. 1994. Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry.* 33:10026–10036.

23. Fersht, A. R. 1995. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl. Acad. Sci. USA.* 92:10869–10873.

24. Fersht, A. R. 1997. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* 7:3–9.

25. Contreras Martínez, L. M., E. E. Borrero Quintana, …, M. P. DeLisa. 2008. In silico protein fragmentation reveals the importance of critical nuclei on domain reassembly. *Biophys. J.* 94:1575–1588.

26. Krantz, B. A., R. S. Dothager, and T. R. Sosnick. 2004. Discerning the structure and energy of multiple transition states in protein folding using psi-analysis. *J. Mol. Biol.* 337:463–475.

27. Levy, Y., P. G. Wolynes, and J. N. Onuchic. 2004. Protein topology determines binding mechanism. *Proc. Natl. Acad. Sci. USA.* 101: 511–516.

28. Dellago, C., and P. G. Bolhuis. 2007. Transition path sampling simulations of biological systems. *In* Atomistic Approaches in Modern Biology: from Quantum Chemistry to Molecular Simulations., *Vol. 268* Springer-Verlag Berlin, Berlin, pp. 291–317.

29. Dellago, C., P. G. Bolhuis, and P. L. Geissler. 2002. Transition path sampling. *Adv. Chem. Phys.* 123:1–78.

30. Borrero, E. E., and F. A. Escobedo. 2006. Folding kinetics of a lattice protein via a forward flux sampling approach. *J. Chem. Phys.* 125:164904–164914.

31. Vendruscolo, M., E. Paci, …, M. Karplus. 2001. Three key residues form a critical contact network in a protein folding transition state. *Nature.* 409:641–645.

32. Borrero, E. E., and F. A. Escobedo. 2007. Reaction coordinates and transition pathways of rare events via forward flux sampling. *J. Chem. Phys.* 127:164101–164117.

33. Borrero, E. E., and F. A. Escobedo. 2008. Optimizing the sampling and staging for simulations of rare events via forward flux sampling schemes. *J. Chem. Phys.* 129:024115–024116.

34. Frenkel, D., and B. Smit. 2002. Understanding Molecular Simulation: From Algorithms to Applications,2nd ed . Academic, Boston, MA.

35. Zhu, W. L., Y. M. Song, …, S. Y. Shin. 2007. Substitution of the leucine zipper sequence in melittin with peptoid residues affects self-association, cell selectivity, and mode of action. *Biochim. Biophys. Acta.* 1768:1506–1517.